Using dbGaP Aggregated Allele Frequency and other large data sets in dbSNP to improve human genetic variation interpretation

Lon Phan- Ph.D.

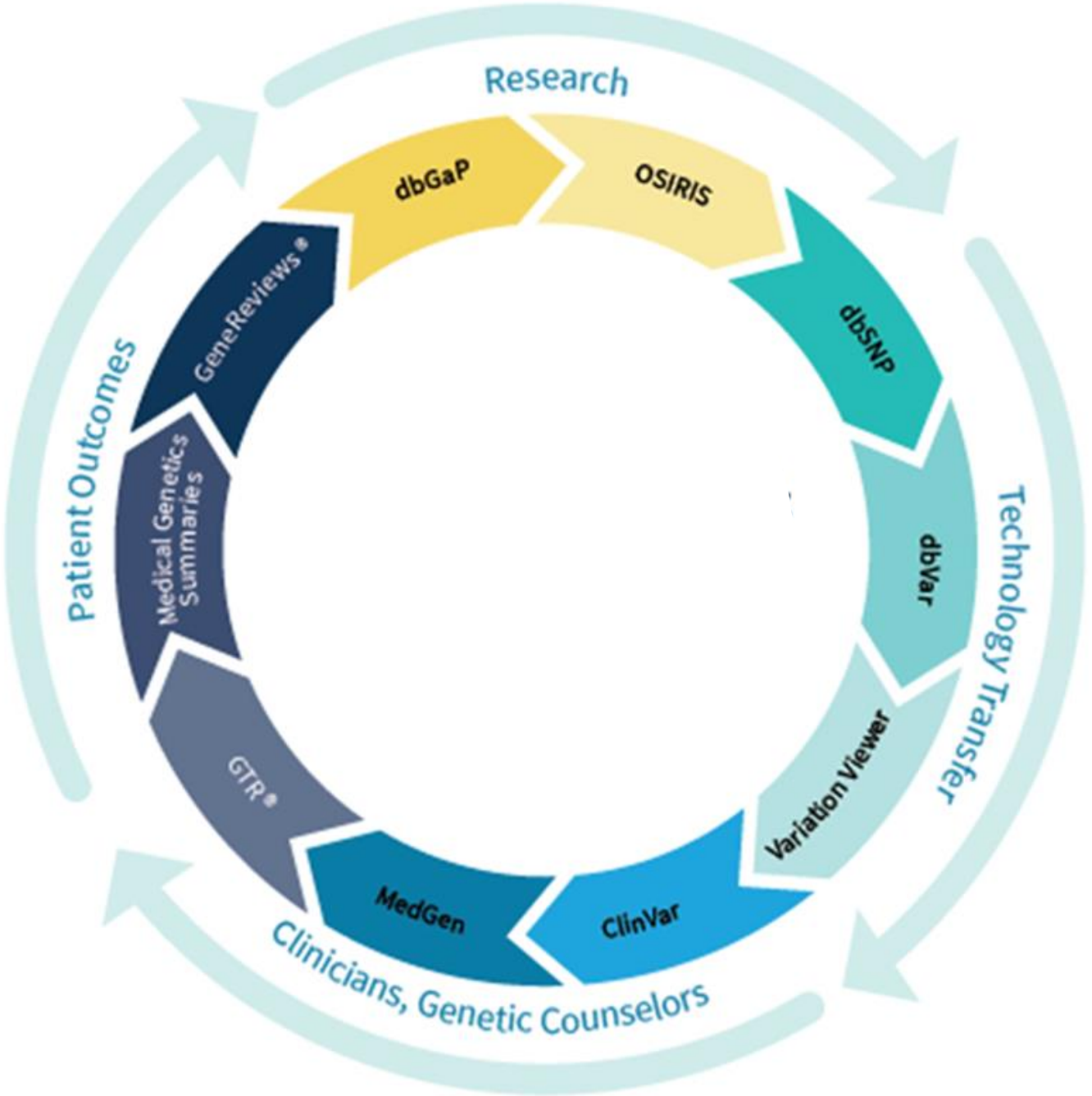
U.S. National Library of Medicine
National Center for Biotechnology Information

Outline

- Introduction
  - dbGaP
  - dbSNP
- Demonstration
  - dbSNP search
  - RefSNP page
  - API and FTP
- Q & A

NCBI Medical Genetics and Human Variation Resources

NCBI Booth:#214



NIH NCBI

dbGaP

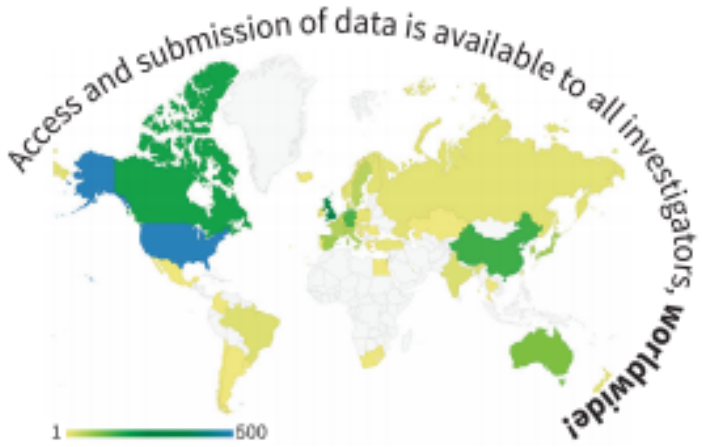
An NIH-sponsored repository charged to archive, curate, and distribute information produced by genome-scale studies investigating the interaction of **human genotype** and **phenotype**.

Your web portal for genotype and phenotype data!

ncbi.nlm.nih.gov/gap/

- **900** released studies
- **Billions** of demograhic, phenotype and exposure measurements
- **1.5 Million** study subjects
- **Trillions** of genotypes
- Over **4500** GWAS analysis datasets
- Over **1,100** publications have referenced use of dbGaP data

- Over **40,000** data access requests from **4,984** investigators in **48** countries.



Contact us at info@ncbi.nlm.nih.gov

Allele Frequency Aggregator (ALFA)

Inputs

Studies	53
Subjects	142,032
Genotypes	696,289,573,125
Genotypes Excluded	791,461,091 (0.1%)

Outputs

RefSNPs	531,167,487
• Exist in dbSNP	512,589,631
• Novel	18,577,856

Coming Soon!!!

NIH NCBI

dbSNP

An archive of

short sequence variants

submitted by the public. dbSNP represents submitted variants, both on the sequences on which each variant was defined, as well as on the current assemblies.

https://www.ncbi.nlm.nih.gov/snp

680 Million Reference SNP (RS) from 2 billion submissions

Mapped to GRCh37 and GRCh38

Allele Frequency for > 550 Million RS

Contact us at info@ncbi.nlm.nih.gov

RefSNP Annotations

- GRCh37 and GRCh38
- RefSeq mRNA and protein
- Functional consequences
- ClinVar Clinical Significance
- Publication
- Allele Frequency
- and many more…

dbSNP Aggregate Frequency Data

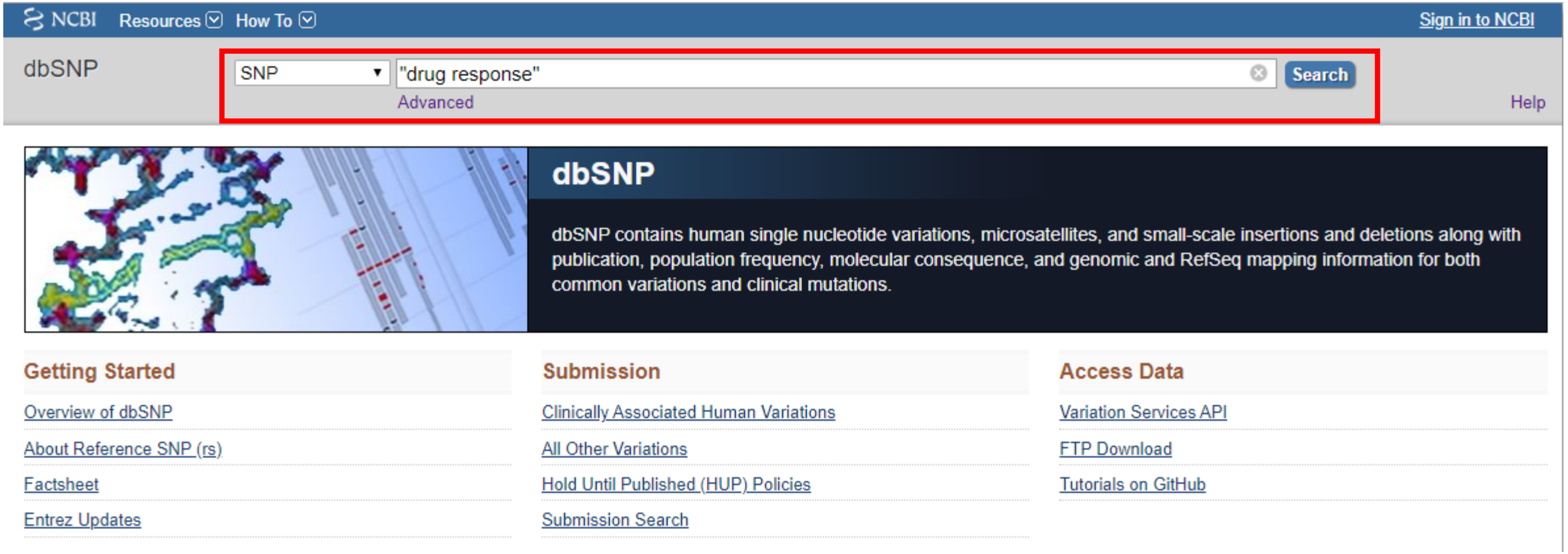
common and rare variants

Diverse Populations		
Project	**Subjects (thousands)**	**Variants (millions)**
ALFA	142.0	531.2
gnomAD	141.5	228.7
TOPMED	62.8	549.4
ExAC	60.7	10.1
PAGE	39.4	1.3
GO-ESP	6.5	1.4
1000 Genomes	2.5	84.9

Regional Populations and Cohorts		
Project	**Subjects (thousands)**	**Variants (millions)**
ALSPAC	3.9	46.6
TWINSUK	3.7	46.6
Estonian	2.2	31.7
Vietnamese	0.3	24.8
Northern Sweden	0.3	17.3

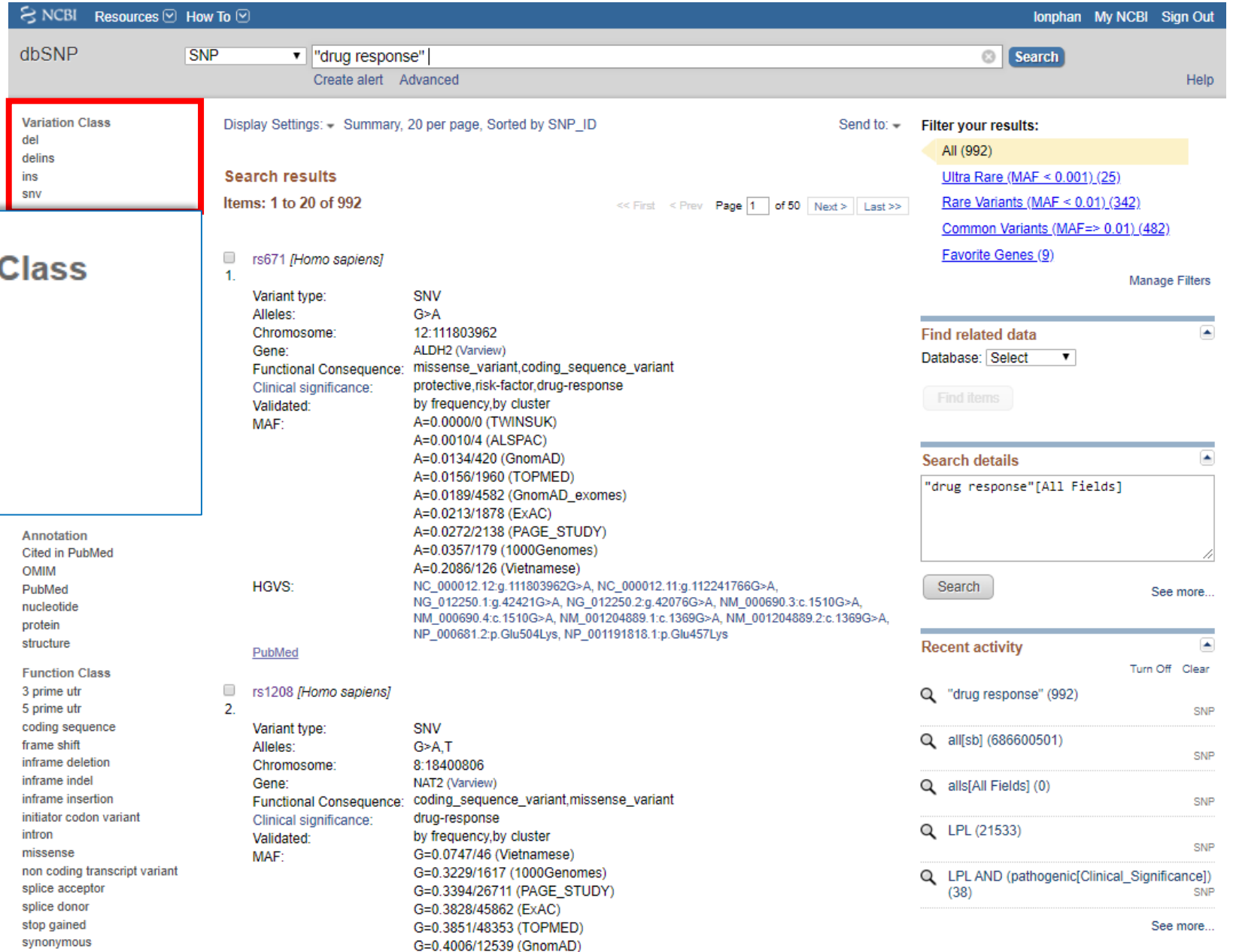
More coming soon!!!

Search dbSNP



https://www.ncbi.nlm.nih.gov/snp

Filter Results

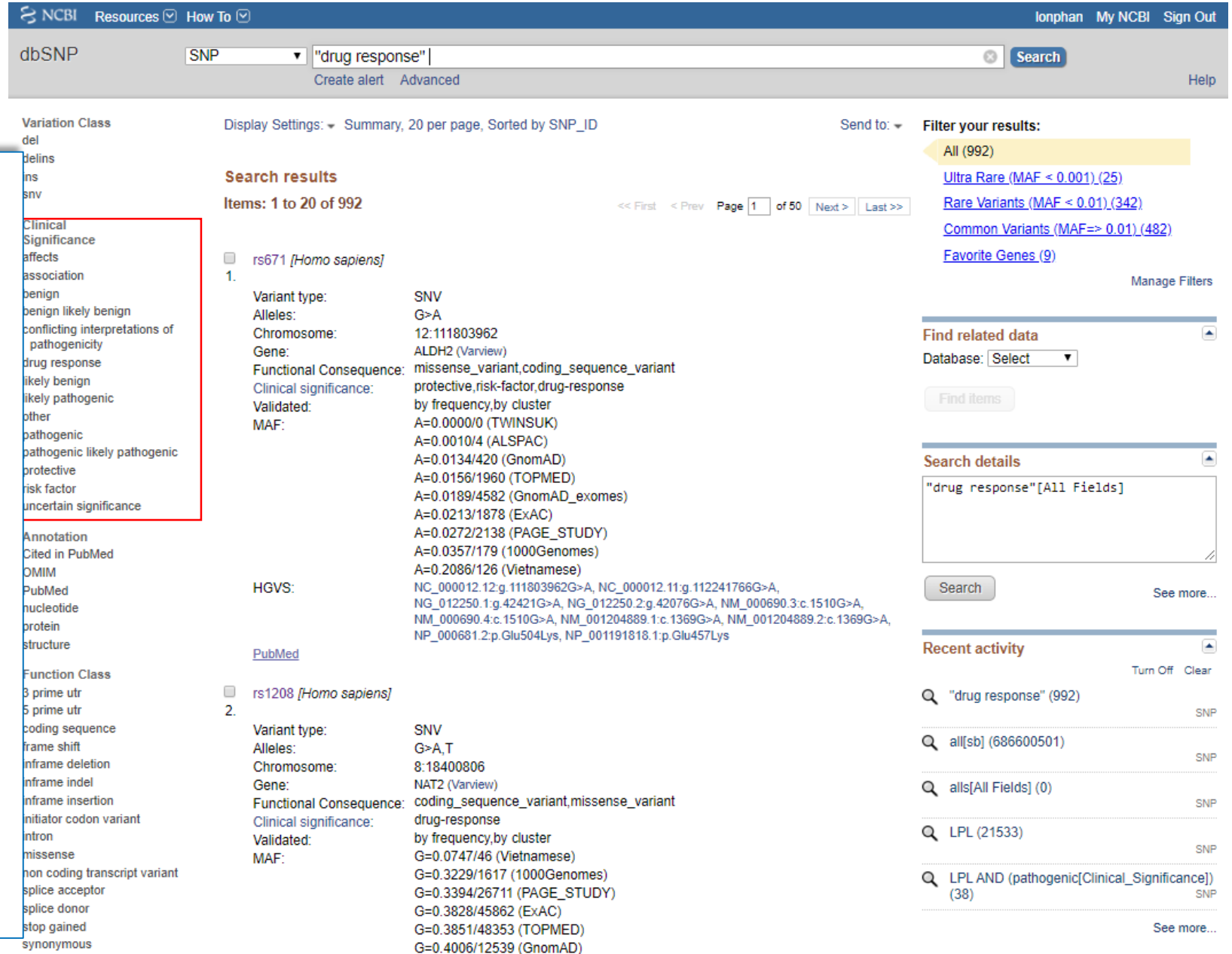


Variation Class
del
delins
ins
snv

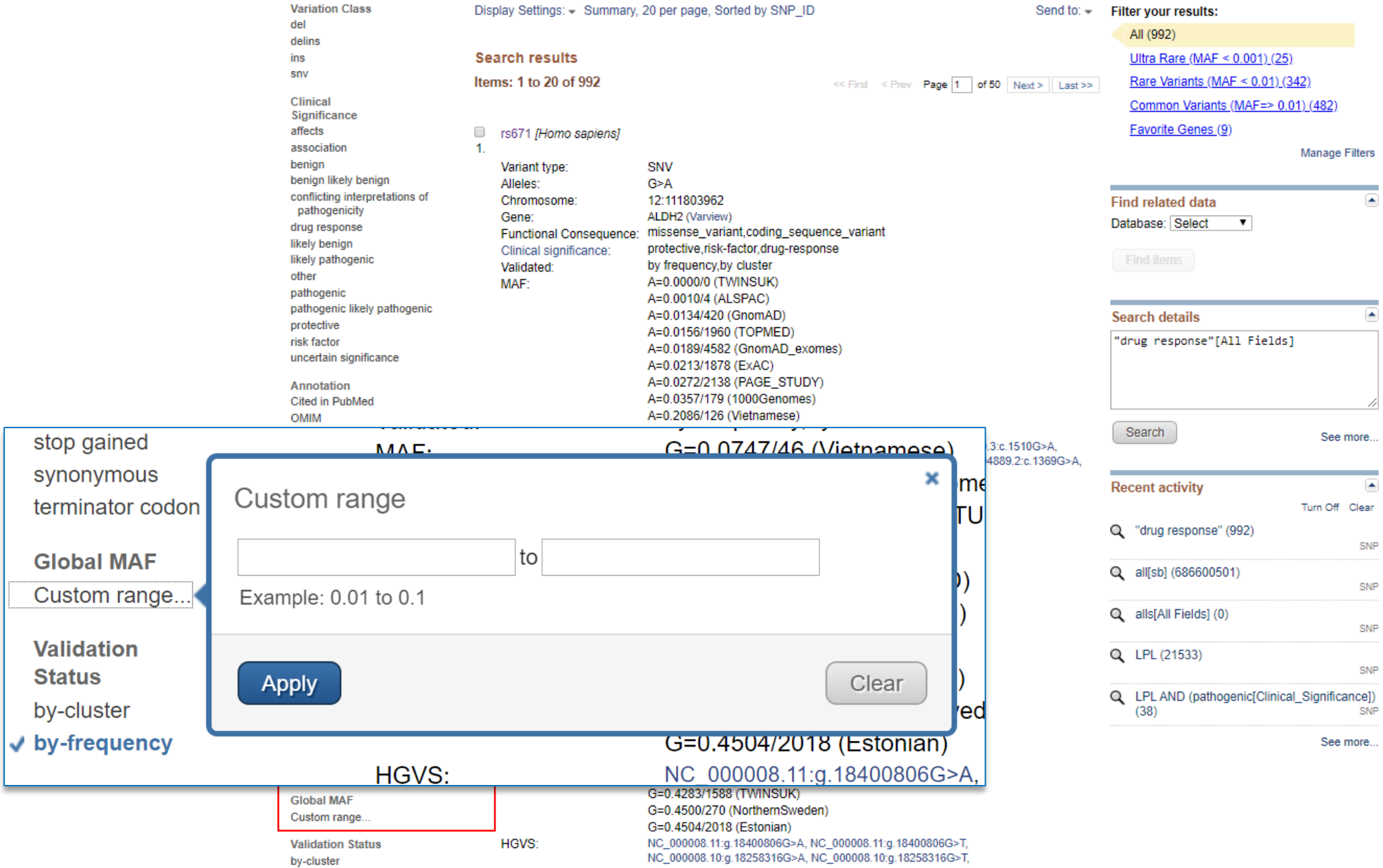
Filter Results

Clinical Significance

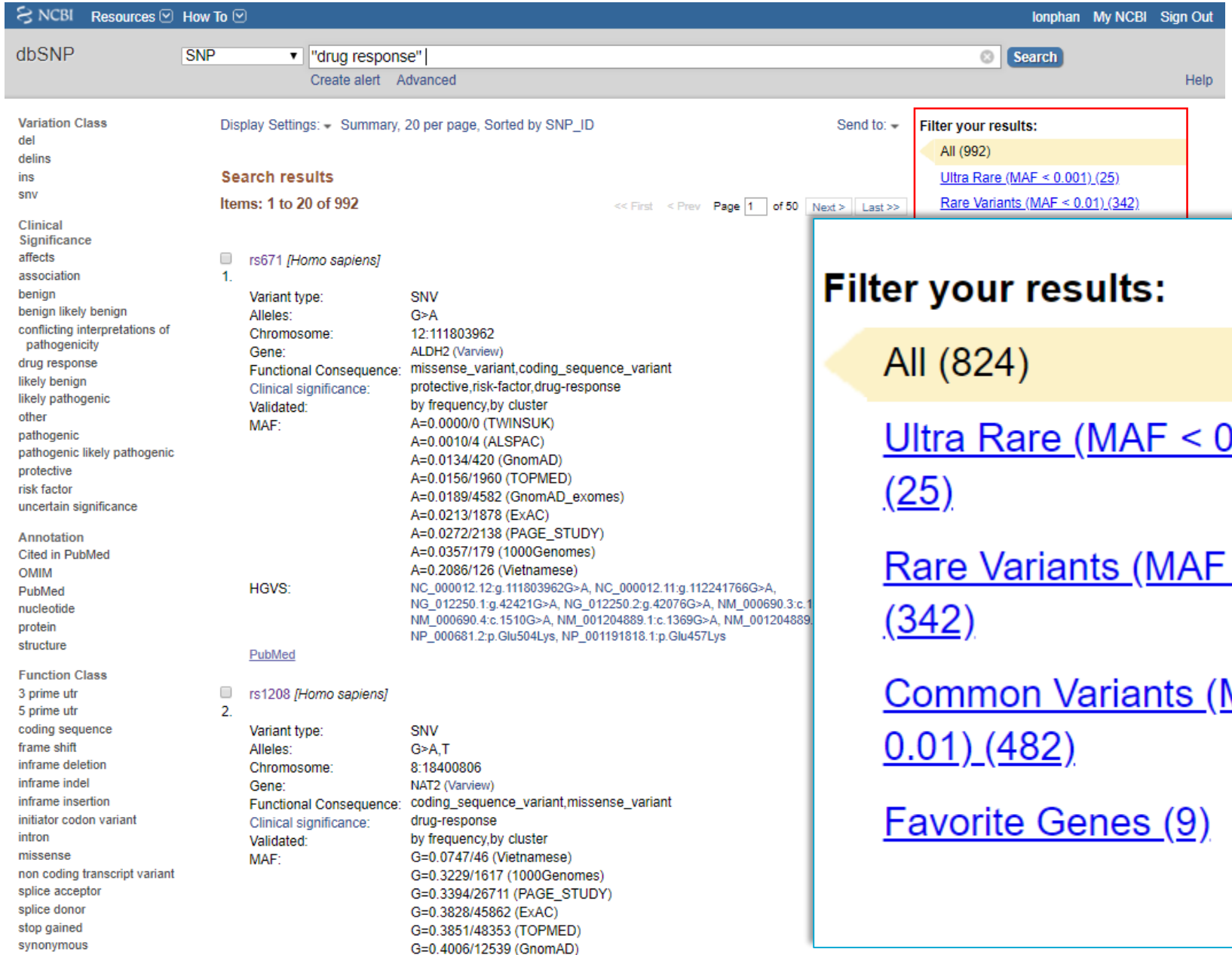
- affects
- association
- benign
- benign likely benign
- conflicting interpretations of pathogenicity
- drug response
- likely benign
- likely pathogenic
- other
- pathogenic
- pathogenic likely pathogenic
- protective
- risk factor
- uncertain significance



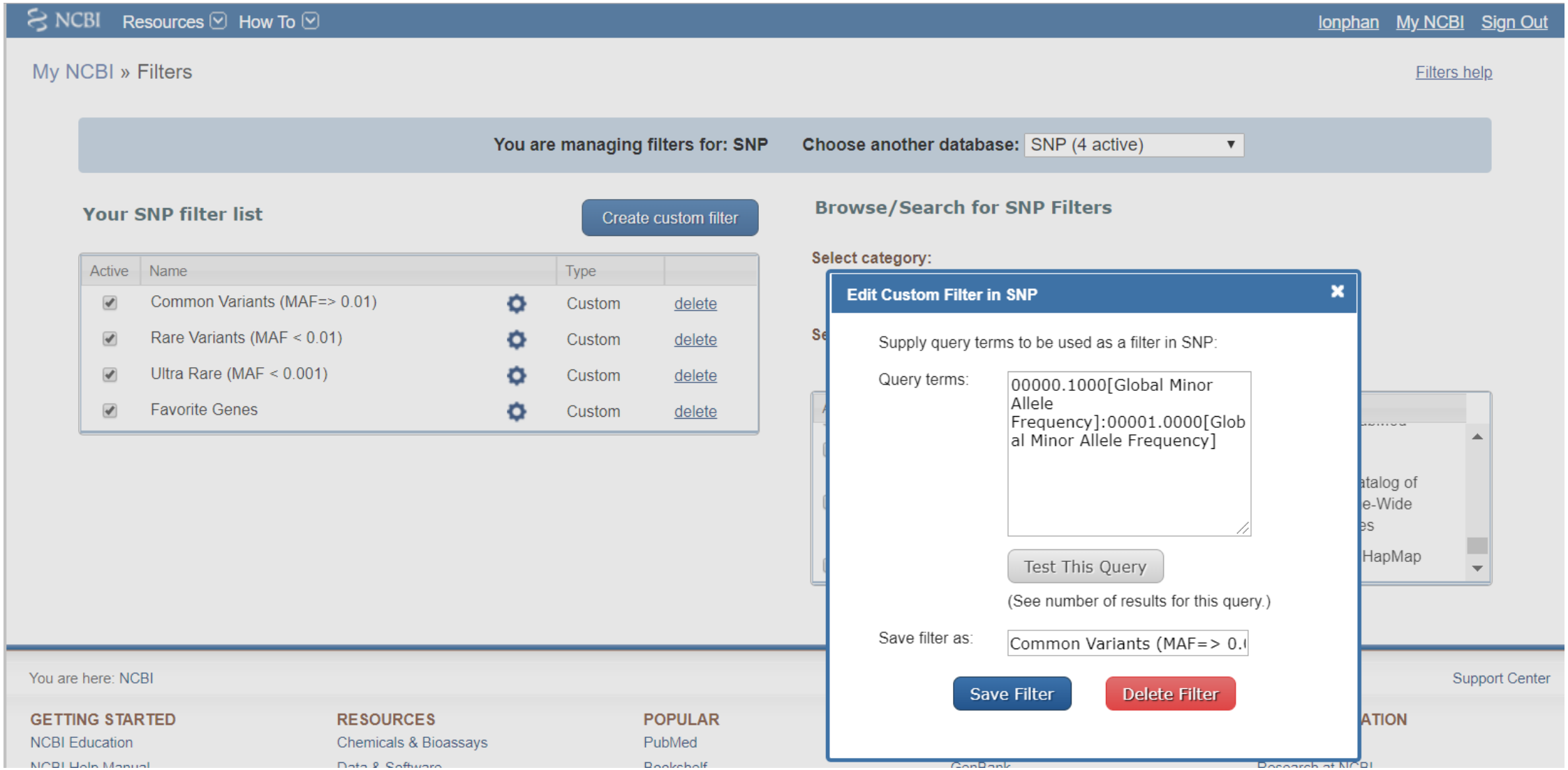
Filter Results



Filter Results



Manage Filters



RefSNP Summary

Search results

Items: 1 to 20 of 992

<< First < Prev Page 1 of 50 Next > Last >>

1. rs671 [Hor

Full RefSNP Report

Variant type:	SNV
Alleles:	G>A
Chromosome:	12:111803962
Gene:	ALDH2 (Varview)
Functional Consequence:	missense_variant,coding_sequence_variant
Clinical significance:	protective,risk-factor,drug-response
Validated:	by frequency,by cluster
MAF:	A=0.0000/0 (TWINSUK) A=0.0010/4 (ALSPAC) A=0.0134/420 (GnomAD) A=0.0156/1960 (TOPMED) A=0.0189/4582 (GnomAD_exomes) A=0.0213/1878 (ExAC) A=0.0272/2138 (PAGE_STUDY) A=0.0357/179 (1000Genomes) A=0.2086/126 (Vietnamese)
HGVS:	NC_000012.12:g.111803962G>A, NC_000012.11:g.112241766G>A, NG_012250.1:g.42421G>A, NG_012250.2:g.42076G>A, NM_000690.3:c.1510G>A, NM_000690.4:c.1510G>A, NM_001204889.1:c.1369G>A, NM_001204889.2:c.1369G>A, NP_000681.2:p.Glu504Lys, NP_001191818.1:p.Glu457Lys

PubMed

RefSNP Report Page

- **Variant Details**
- Clinical Significance
- Frequency
- Aliases
- Submissions
- History
- Publications

U.S. National Library of Medicine
National Center for Biotechnology Information

lonphan

bSNP Short Genetic Variations

Search for rs — Search
Example: rs268

ference SNP (rs) Report

Download

witch to classic site

rs671

Current Build 153
Released July 9, 2019

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr12:111803962 (GRCh38.p12)	**Gene : Consequence**	ALDH2 : Missense Variant
Alleles	G>A	**Publications**	193 citations
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome
Frequency	A=0.01888 (4582/242666, GnomAD_exome) A=0.01561 (1960/125568, TOPMED) A=0.0213 (1878/88224, ExAC) (+ 6 more)		

- Variant Details
- Clinical Significance
- Frequency
- Aliases
- Submissions
- History
- Publications

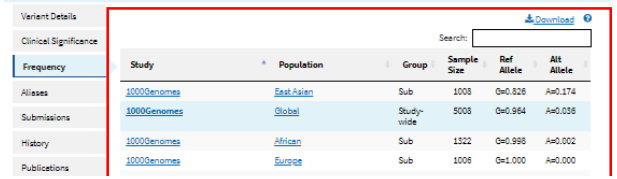
Genomic Placements

Sequence name	Change
ALDH2 RefSeqGene	NG_012250.1:g.42421G>A
ALDH2 RefSeqGene	NG_012250.2:g.42076G>A
GRCh37.p13 chr 12	NC_000012.11:g.112241766G>A
GRCh38.p12 chr 12	NC_000012.12:g.111803962G>A

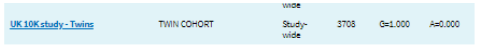
Gene: ALDH2, aldehyde dehydrogenase 2 family member (plus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
aldehyde dehydrogenase, mitochondrial isoform 1 precursor	NP_000681.2:p.Glu504Lys	E (Glu) > K (Lys)	Missense Variant

RefSNP Allele Frequency

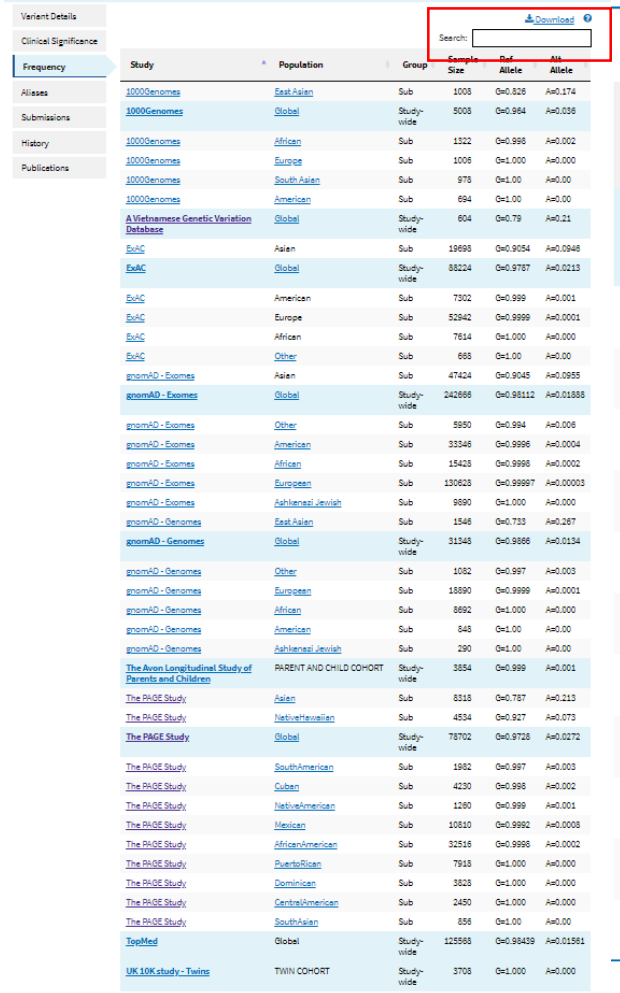


Study	Population	Group	Sample Size	Ref Allele	Alt Allele
1000Genomes	Global	Study-wide	5008	G=0.964	A=0.036
1000Genomes	African	Sub	1322	G=0.998	A=0.002
1000Genomes	East Asian	Sub	1008	G=0.826	A=0.174
1000Genomes	Europe	Sub	1006	G=1.000	A=0.000
1000Genomes	South Asian	Sub	978	G=1.00	A=0.00
1000Genomes	American	Sub	694	G=1.00	A=0.00



RefSNP Allele Frequency

Filter Using Search Box



Search: PAGE

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
The PAGE Study	Global	Study-wide	78702	G=0.9728	A=0.0272
The PAGE Study	AfricanAmerican	Sub	32516	G=0.9998	A=0.0002
The PAGE Study	Mexican	Sub	10810	G=0.9992	A=0.0008
The PAGE Study	Asian	Sub	8318	G=0.787	A=0.213
The PAGE Study	PuertoRican	Sub	7918	G=1.000	A=0.000
The PAGE Study	NativeHawaiian	Sub	4534	G=0.927	A=0.073
The PAGE Study	Cuban	Sub	4230	G=0.998	A=0.002
The PAGE Study	Dominican	Sub	3828	G=1.000	A=0.000
The PAGE Study	CentralAmerican	Sub	2450	G=1.000	A=0.000
The PAGE Study	SouthAmerican	Sub	1982	G=0.997	A=0.003
The PAGE Study	NativeAmerican	Sub	1260	G=0.999	A=0.001
The PAGE Study	SouthAsian	Sub	856	G=1.00	A=0.00

RefSNP Allele Frequency

Sortable Columns

Study	Population	Group	Sample Size	Ref Allele	Alt Allele
gnomAD - Genomes	East Asian	Sub	1546	G=0.733	A=0.267
The PAGE Study	Asian	Sub	8318	G=0.787	A=0.213
A Vietnamese Genetic Variation Database	Global	Study-wide	604	G=0.79	A=0.21
1000Genomes	East Asian	Sub	1008	G=0.826	A=0.174
gnomAD - Exomes	Asian	Sub	47424	G=0.9045	A=0.0955
ExAC	Asian	Sub	19698	G=0.9054	A=0.0946
The PAGE Study	NativeHawaiian	Sub	4534	G=0.927	A=0.073
1000Genomes	Global	Study-wide	5008	G=0.964	A=0.036
The PAGE Study	Global	Study-wide	78702	G=0.9728	A=0.0272
ExAC	Global	Study-wide	88224	G=0.9787	A=0.0213
gnomAD - Exomes	Global	Study-wide	242666	G=0.98112	A=0.01888

ALFA frequency reported on RefSNP Page

Variant Details
Clinical Significance
Frequency
Aliases
Submissions
History
Publications

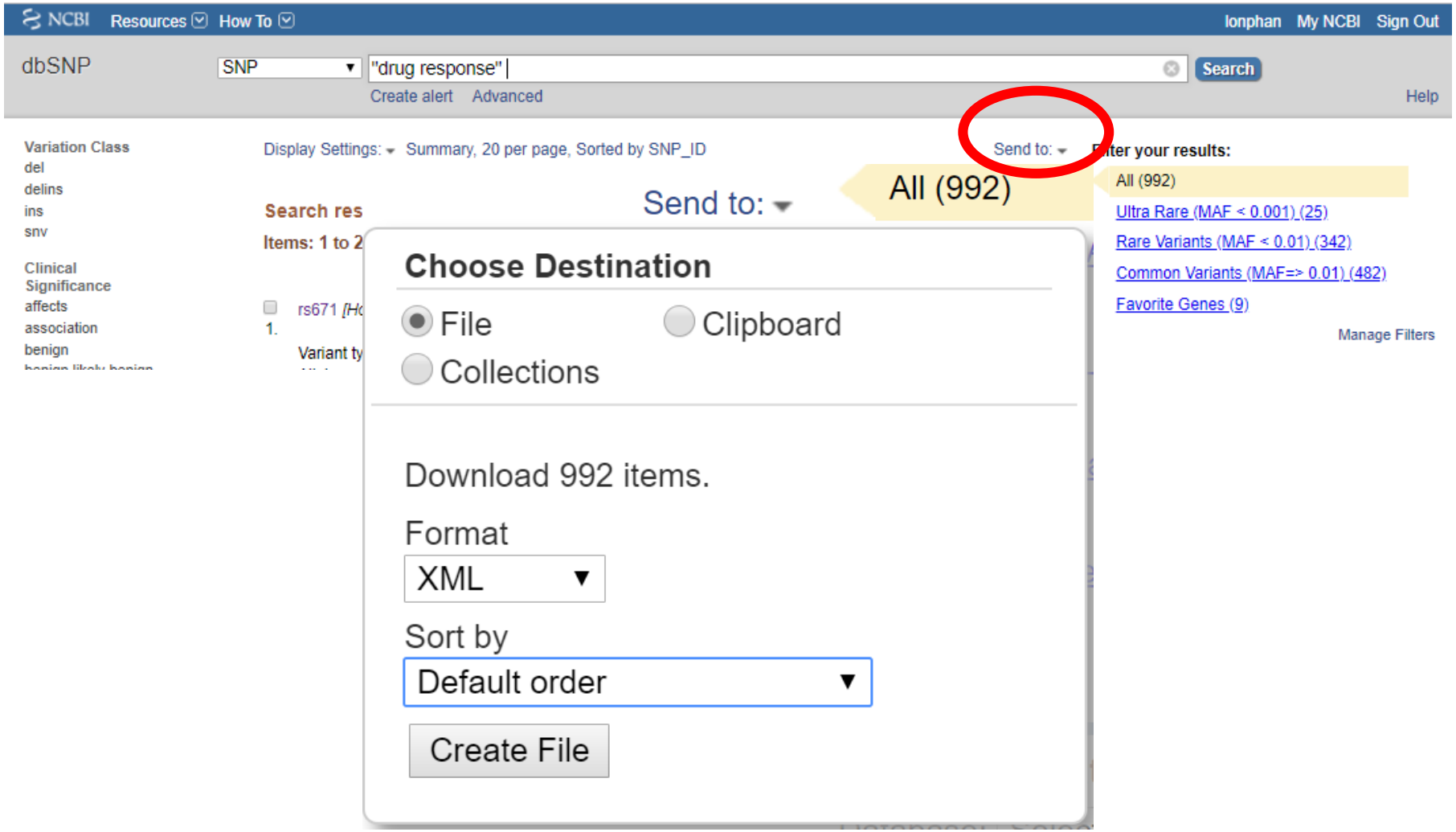
dbGaP Population Frequency Project **Release Version:** 20190529232800

Population	Group	Sample Size	Ref Allele	Alt Allele
Global	Global	173172	T=0.43013	C=0.56987
Europe	Sub	145214	T=0.42688	C=0.57312
All African Ancestry	Sub	8664	T=0.479	C=0.521
95% Exclusive African Ancestry	Sub	304	T=0.48	C=0.52
African American	Sub	8360	T=0.479	C=0.521
Asian	Sub	4546	T=0.506	C=0.494
95% East Asian Ancestry	Sub	4264	T=0.507	C=0.493
South East Asian and Pacific Islanders	Sub	282	T=0.48	C=0.52
Latin American 1	Sub	1032	T=0.390	C=0.610
Latin American 2	Sub	2134	T=0.344	C=0.656
South Asian	Sub	5020	T=0.438	C=0.562
Other	Sub	6562	T=0.413	C=0.587

Filter: Download

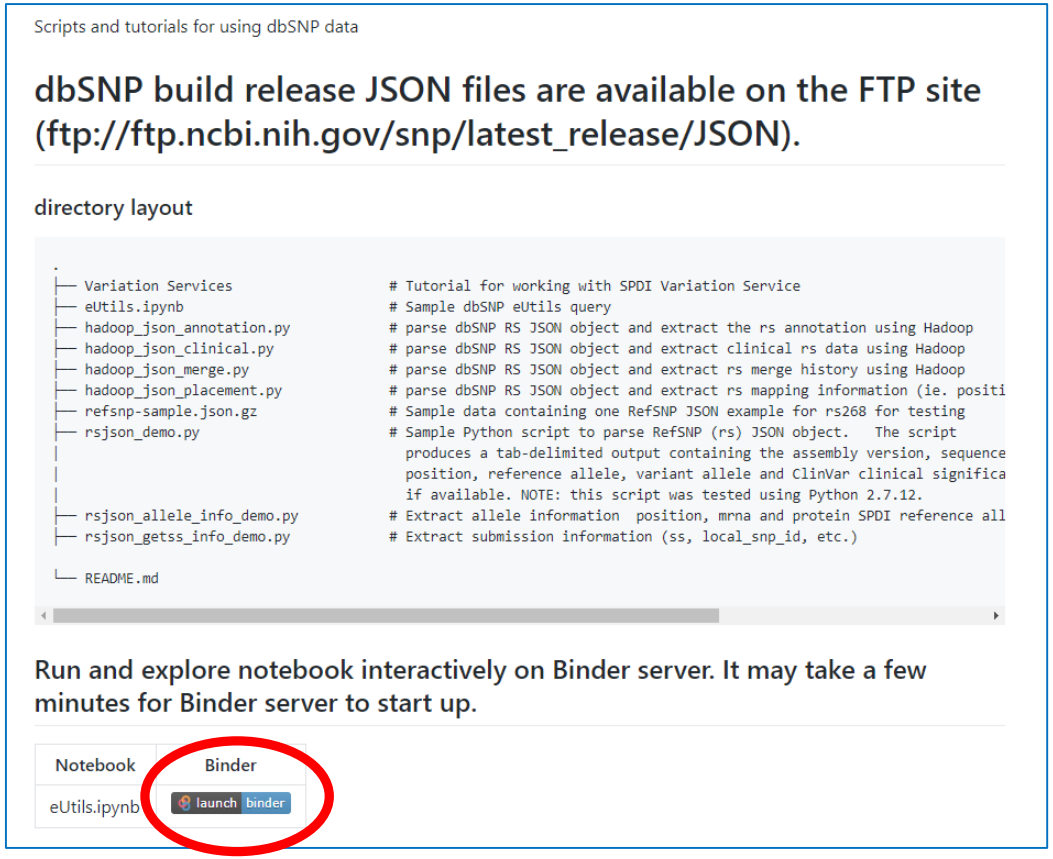
Study	Population	Group	Sample Size	Ref Allele	Alt Allele
TopMed	Global	Study-wide	125568	T=0.43702	C=0.56298
The PAGE Study	Global	Study-wide	78702	T=0.4254	C=0.5746
The PAGE Study	AfricanAmerican	Sub	32516	T=0.4690	C=0.5310
The PAGE Study	Mexican	Sub	10810	T=0.3455	C=0.6545

Save and Download Results



Search and Retrieve Using eUtils

https://github.com/ncbi/dbsnp



Scripts and tutorials for using dbSNP data

dbSNP build release JSON files are available on the FTP site (ftp://ftp.ncbi.nih.gov/snp/latest_release/JSON).

directory layout

```
.
├── Variation Services                # Tutorial for working with SPDI Variation Service
├── eUtils.ipynb                      # Sample dbSNP eUtils query
├── hadoop_json_annotation.py         # parse dbSNP RS JSON object and extract the rs annotation using Hadoop
├── hadoop_json_clinical.py           # parse dbSNP RS JSON object and extract clinical rs data using Hadoop
├── hadoop_json_merge.py              # parse dbSNP RS JSON object and extract rs merge history using Hadoop
├── hadoop_json_placement.py          # parse dbSNP RS JSON object and extract rs mapping information (ie. positi
├── refsnp-sample.json.gz             # Sample data containing one RefSNP JSON example for rs268 for testing
├── rsjson_demo.py                    # Sample Python script to parse RefSNP (rs) JSON object.   The script
│                                       produces a tab-delimited output containing the assembly version, sequence
│                                       position, reference allele, variant allele and ClinVar clinical significa
│                                       if available. NOTE: this script was tested using Python 2.7.12.
├── rsjson_allele_info_demo.py        # Extract allele information  position, mrna and protein SPDI reference all
├── rsjson_getss_info_demo.py         # Extract submission information (ss, local_snp_id, etc.)
│
└── README.md
```

Run and explore notebook interactively on Binder server. It may take a few minutes for Binder server to start up.

Notebook	Binder
eUtils.ipynb	launch binder

eUtils Jupyter Notebook

jupyter eUtils (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Code

In [2]:
```
Entrez.email = "dbsnp-user@nih.gov" # provide your user email
# RECOMMENDED: apply for API key from NCBI (https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/).
# 10 queries per second with a valid API key, otherwise 3 queries per seconds are allowed for 'None'
Entrez.api_key = None

# entrez query (term) can be build and test online using web query builder (https://www.ncbi.nlm.nih.gov/snp/advanced)
eShandle = Entrez.esearch(db="snp",  # search dbSNP
                          #complex query for missense and pathogenic variants in LPL gene with global MAF betweeen 0 and 0.01.
                          term='LPL[All Fields] AND pathogenic[Clinical_Significance] AND missense variant[Function_Class] AND (
                          usehistory="y", #cache result on server for download in batches
                          retmax=20 # return 20 RSID max
```


```
eShandle = Entrez.esearch(db="snp",  # search dbSNP
                          #complex query for missense and pathogenic variants in LPL gene with global MAF betweeen 0
                          term='LPL[All Fields] AND pathogenic[Clinical_Significance] AND missense variant[Function_
                          usehistory="y", #cache result on server for download in batches
                          retmax=20 # return 20 RSID max
                         )
```


```
#https://www.ncbi.nlm.nih.gov/books/NBK25500/#chapter1.Storing_Search_Results
```

```
Count : 5
RetMax : 5
RetStart : 0
QueryKey : 1
WebEnv : NCID_1_3075896_130.14.18.97_9001_1567108675_833541832_0MetA0_S_MegaStore
IdList : ['386571803', '118204057', '52818902', '17850737', '268']
```

Summary

- dbSNP
  - 680 Million Reference SNP (RS)
  - 550 Million RS with frequency aggregated from 1000Genomes, GnomAD, TopMed, and others
- dbGaP (ALFA) has the largest dataset will be coming soon.
- Robust search and retrieval systems (web and API)

Link to the presentation will be available after ASHG

https://www.ncbi.nlm.nih.gov/snp

Get Announcements, Updates, Releases, and More
Subscribe to dbSNP today!

https://go.usa.gov/xVdJg

NCBI at ASHG 2019 booth #214!